

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平4-233025

(43) 公開日 平成4年(1992)8月21日

(51) Int.Cl. ⁵	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	3 0 5 F	7165-5B		
11/10	3 3 0 L	9072-5B		

審査請求 有 請求項の数13(全 10 頁)

(21) 出願番号 特願平3-77305

(22) 出願日 平成3年(1991)3月18日

(31) 優先権主張番号 5 4 2 2 1 6

(32) 優先日 1990年6月21日

(33) 優先権主張国 米国 (U S)

(71) 出願人 390009531

インターナショナル・ビジネス・マシーンズ・コーポレーション

INTERNATIONAL BUSINESS MACHINES CORPORATION

アメリカ合衆国10504、ニューヨーク州アーモンク (番地なし)

(72) 発明者 ミルトン・フレドリック・ボンド

アメリカ合衆国 55901、ミネソタ州、ロチェスター、16 1/2 アベニュー・ノース・ウエスト 1520番地

(74) 代理人 弁理士 頓宮 孝一 (外4名)

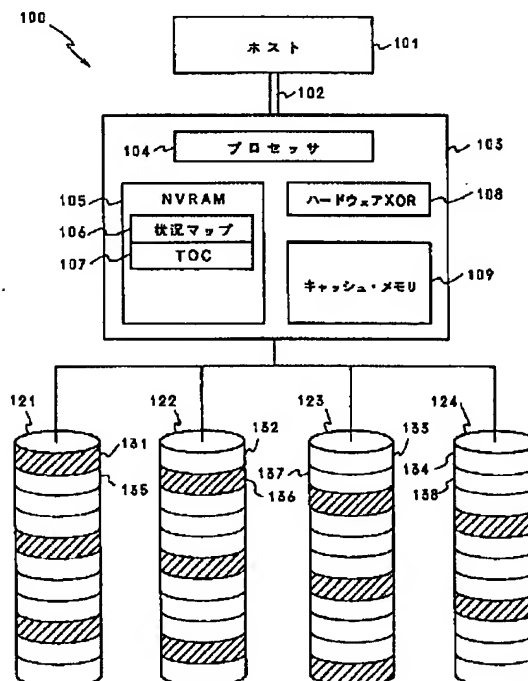
最終頁に続く

(54) 【発明の名称】 パリティ保護データを回復するための方法および装置

(57) 【要約】

【目的】 複数のデータ記憶装置を有するコンピュータ・システムにおけるデータ・エラーからの回復方法および装置を提供する。

【構成】 記憶管理機構が記憶装置のパリティ・レコードを維持し、データ・ブロックごとにその位置と状況を示すマップを持つ。エラー時にマップを検査し、データの再構築が未了の場合はすべての記憶装置の該当ブロックの排他ORを累計してデータの再構築をして、ブロック再構築済みであることを示すように更新される。その後のアクセスは以前のパリティ・ブロックを用いて実行される。故障していない記憶装置内の領域を再構築の際に割り振ることも可能である。



(2)

特開平4-233025

1

【特許請求の範囲】

【請求項1】データを収容するための複数のデータ記憶ブロックと、上記データ記憶ブロックに記憶されたデータのパリティを収容するための1つのパリティ記憶ブロックとを含む、記憶ブロックのストライプを有し、上記の各記憶ブロックがそれぞれ当該のデータ記憶装置上に含まれる、コンピュータ・システムを動作させる方法であって、ある記憶ブロックを含むデータ記憶装置が故障しているとき、ストライプ中の残りの記憶ブロックから、上記の記憶ブロックに含まれるデータを再構築するステップと、上記再構築ステップによって再構築されたデータを、上記データ記憶装置の1台に記憶するステップとを含む方法。

【請求項2】上記のデータ再構築ステップが、あるデータにアクセスが試みられたとき、そのデータを再構築することを特徴とする、請求項1に記載のコンピュータ・システムを動作させる方法。

【請求項3】上記のデータ記憶ステップが、再構築されたデータを上記のパリティ記憶ブロックに記憶することを特徴とする、請求項1に記載のコンピュータ・システムを動作させる方法。

【請求項4】上記のデータ再構築ステップが、あるデータにアクセスが試みられたとき、そのデータを再構築することを特徴とする、請求項3に記載のコンピュータ・システムを動作させる方法。

【請求項5】上記データ記憶装置が予備記憶ブロックを含み、上記のデータ記憶ステップが、再構築されたデータを上記の予備記憶ブロックに記憶することを特徴とする、請求項1に記載のコンピュータ・システムを動作させる方法。

【請求項6】上記のデータ再構築ステップが、あるデータにアクセスが試みられたとき、そのデータを再構築することを特徴とする、請求項5に記載のコンピュータ・システムを動作させる方法。

【請求項7】少なくとも3台のデータ記憶装置と、各ストライプが、データを収容するための複数のデータ記憶ブロックと、上記データ記憶ブロックに記憶されたデータのパリティを収容するための1つのパリティ記憶ブロックとを含む、上記の各記憶ブロックがそれぞれ当該のデータ記憶装置上に含まれる、記憶ブロックの少なくとも1つのストライプと、ある記憶ブロックを含むデータ記憶装置が故障しているとき、ストライプ中の残りの記憶ブロックから、上記の記憶ブロックに含まれるデータを再構築する手段と、上記の再構築されたデータを上記のデータ記憶装置の1つに記憶する手段とを含む、コンピュータ・システム用の記憶装置。

【請求項8】上記の再構築データ記憶手段が、上記のデータを上記のパリティ記憶ブロックに記憶することを特徴とする、請求項7に記載のコンピュータ・システム用の記憶装置。

2

【請求項9】上記のデータ再構築手段が記憶制御装置を含み、上記記憶制御装置が、記憶管理プログラムを実行するプログラム式プロセッサと、持久ランダム・アクセス記憶装置とを含むことを特徴とする、請求項7に記載のコンピュータ・システム用の記憶装置。

【請求項10】上記のデータ処理システムが、上記の記憶ブロックのストライプを少なくとも2本含み、かつ上記のパリティ記憶ブロックが、上記のデータ記憶装置の間にラウンド・ロビン方式で分配されることを特徴とする、請求項8に記載のコンピュータ・システム用の記憶装置。

【請求項11】上記の各データ記憶装置が、回転式磁気ディスク・ドライブ記憶装置であることを特徴とする、請求項8に記載のコンピュータ・システム用の記憶装置。

【請求項12】上記の各データ記憶装置が予備記憶ブロックを含み、かつ上記の再構築データ記憶手段が、上記のデータを上記の予備記憶ブロックの1つに記憶することを特徴とする、請求項7に記載のコンピュータ・システム用の記憶装置。

【請求項13】上記の各データ記憶装置が、回転式磁気ディスク・ドライブ記憶装置であることを特徴とする、請求項12に記載のコンピュータ・システム用の記憶装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、コンピュータのデータ記憶装置に関するパリティ情報の維持に関し、より具体的には、故障した記憶装置からのデータを再構築する際に、コンピュータ・システムの使用可能性を維持することに関する。

【0002】

【従来の技術】現代のコンピュータ・システムのデータ記憶ニーズは、大容量のデータ記憶装置を必要としている。普通に使われている記憶装置は磁気ディスク装置であり、故障しやすい多数の部品を含む複雑な機械である。通常のコンピュータ・システムは、この磁気ディスク装置を複数台含んでいる。ユーザがそのデータ記憶ニーズを増大させるにつれて、より多くの記憶装置を含むようにシステムが構成されていく。このシステムにとって、一台の記憶装置の故障が、非常に破壊的な事象となる可能性がある。多くのシステムは、欠陥装置が修理または交換され、失われたデータが復元されるまで、動作できない。記憶装置の台数が増えるにつれて、いずれか1台の装置が故障して、システム障害をもたらす確率も増大する。同時に、コンピュータ・ユーザは、ますますそのシステムの一貫した使用可能性に対する依存を深めている。したがって、故障した記憶装置に含まれていたデータを再構築し、記憶装置の故障が存在する状態でシステムの動作を持続させる、改良された方法を見つける

(3)

特開平4-233025

3

ことが必須になってきた。

【0003】これらの問題に対処する一つの方法は、「ミラー（対称）構成」と呼ばれるものである。この方法は、元のデータと同じデータを含む、重複した1セットの記憶装置を維持するものである。元のセット中のいずれかの装置が故障した場合、この重複セットを使って、システムにデータを供給するタスクを引き受けることができる。これは、この問題を解決するための非常に有効な方法であるが、顧客が2倍の記憶装置の代金を支払わなければならないので、非常に高くつく。

【0004】より安価な別の方法は、パリティ・ブロックを使用するものである。パリティ・ブロックとは、異なる記憶装置上の特定の位置に記憶されたすべてのデータ・レコードの排他的ORによって形成されるレコードである。言い換えれば、記憶装置の特定の位置にあるデータ・ブロック中の各ビットが、装置グループ中の各記憶装置の同じ位置にある他のあらゆるビットと排他的ORされて、パリティ・ビットのブロックを生成し、このパリティ・ブロックが別の記憶装置の同じ位置に記憶される。そのグループ中のいずれかの記憶装置が故障した場合、残りの装置の同じ位置にあるデータ・ブロックとそれに対応するパリティ・ブロックの排他的ORをとることにより、故障した装置のいずれかの位置に含まれるデータを再生させることができる。

【0005】米国特許第4092732号は、パリティ・ブロック法を記載している。この特許の装置では、1台の記憶装置を使って、1群の記憶装置のパリティ情報を記憶する。そのパリティ・レコードでカバーされるグループ中のいずれかの記憶装置上でレコードが変更されるごとに、パリティ・ブロックを含む記憶装置での読取りと書き込みが行なわれる。したがって、パリティ・レコードを含む記憶装置が、記憶動作のネックとなる。米国特許第4761785号明細書は、パリティ・ブロックを1セットの記憶装置の間にほぼ均等に分配することにより、パリティ情報の記憶を改善するものである。この特許を、引用により本明細書に合体する。1セット中のN個の記憶装置が、それぞれが複数のレコードを含む、サイズの等しい多数のアドレス・ブロックに分割される。同じアドレス範囲をもつ各記憶装置からのブロックが集まって、ブロック・ストライプを形成する。各ストライプは、1台の記憶装置上にそのストライプの残りのブロックのパリティを含むブロックを有する。異なるストライプのパリティ・ブロックが、異なる記憶装置の間にラウンド・ロビン方式で分配される。

【0006】前記の米国特許第4092732号および第4761785号明細書に記載されているようにパリティ・レコードを使用すると、ミラー構成に比べてデータ保護のコストが大幅に低下する。しかし、上記の両特許はデータの回復または保護手段は教示しているが、データの再構築中にシステムを動作状態に保つ手段は提供

4

していない。故障した記憶装置を修理または交換し、続いてデータを再構築するためにメモリ制御装置を停止する間、正常な動作は中断される。この従来技術は専らデータの再構築に依拠しているため、システムがかなりの時間動作不能になることがある。

【0007】従来技術は、重複したまたは待機の記憶装置を使用しない動的システム回復および継続動作を教示していない。ミラー構成では、記憶装置の数を倍にする必要がある。それほど極端でない手法は、1台以上の待機装置、すなわち元来のセット中のいずれかの装置が故障した場合にオンラインにすることのできる、追加の予備ディスク・ドライブを使用するものである。この手法は、完全にミラー構成のシステムほどコストはかからないが、それでも通常は有用な機能を果さない、追加の記憶装置が必要である。

【0008】

【発明が解決しようとする課題】本発明の一目的は、多数のデータ記憶装置をもつコンピュータ・システムでデータ喪失から回復するための改良された方法および装置を提供することにある。

【0009】本発明の他の目的は、複数のデータ記憶装置をもつコンピュータ・システムが、データ記憶装置が1台故障しても引き続き動作できる、改良された方法および装置を提供することにある。

【0010】本発明の他の目的は、複数の保護された記憶装置をもつデータ処理システムにおけるデータ保護のコストを節減することにある。

【0011】本発明の他の目的は、データ記憶装置の1台が故障して、システムが故障した装置に含まれるデータを再構築しなければならないときに、複数のデータ記憶装置をもつコンピュータ・システムの性能を向上させることにある。

【0012】

【課題を解決するための手段】記憶制御装置は、複数のデータ記憶装置にサービスする。制御装置上にある記憶管理機構が、そのサービスする記憶装置上のパリティ・レコードを維持する。データとパリティ・ブロックは、上記の米国特許第4761785号に記載されているように編成する。記憶装置が故障した場合でも、システムは動作し続ける。記憶管理機構は、故障装置へのアクセスが試みられたとき、その装置上にあったデータを再構築し、それを残りの記憶装置のパリティ・ブロック域に記憶する。

【0013】記憶管理機構は、各データ・ブロックごとに、対応するパリティ・ブロックの位置とデータ・ブロックの状況とを示す状況マップを含む。ある記憶装置が故障した場合、記憶管理機構は故障動作モードになる。故障動作モードの間、記憶管理機構は、故障した記憶装置にアクセスする前に、状況マップを検査する。データがまだ再構築されていない場合、記憶管理機構はまず、

(4)

特開平4-233025

5

そのパリティ・ブロックを含むパリティ・グループ内の、すべての記憶装置上の同じブロックの排他的OR (XOR) を次々に読み取って累計することにより、その記憶ブロック内のデータを再構築しなければならない。この排他的ORの結果得られるデータ・ブロックが再構築されたデータであり、次にそれがそのパリティ・ブロックの位置に記憶される。次いで状況マップが、そのブロックが再構築されたことを示すように更新される。データが再構築された後は、以前のパリティ・ブロックから直接読み取る、または以前のパリティ・ブロックに直接書き込むだけでよい。同様に、記憶管理機構は、(故障していない装置上の) 同じストライプ上の他のどのブロックに書き込む前にも、故障した装置上の記憶ブロックからデータを再構築する。このことが必要なのは、そのストライプ上のいずれかのブロックへの書き込み動作でパリティが変化し、後で故障した装置上のデータ・ブロックを再構築することが不可能になるからである。すなわち、ある記憶装置が故障すると、読取り動作および書き込み動作の際に記憶管理機構がデータを再構築するので、最初はシステム性能が低下する。データが再構築された後は、機能は速やかに向上する。

【0014】好ましい実施例では、記憶装置の編成ならびにパリティ情報の生成と記憶は、前記の米国特許第4761785号明細書に記載の通りである。再構築されたデータは、失われたデータが存在したストライプに対するパリティ・データが通常なら記憶されるはずの場所に記憶される。記憶制御装置またはシステムの他のいずれかの部分を停止し、故障した記憶装置を修理し、失われたデータを再構築する必要がある。この好ましい実施例では、この記憶管理機構がユーザに完全に利用可能なまま、データが回復され記憶される。記憶装置は、故障した装置が修理または変換されるまで、パリティ保護なしで動作する。この実施例は、非常にわずかな追加コストで、連続運転と単一レベルの障害保護を実現する。

【0015】第1の代替実施例では、再構築されたデータに、故障していない各記憶装置の予備記憶域が割り当てられる。これらの予備記憶域の全体が、仮想予備記憶装置を構成する。データが再構築されたとき、それは仮想予備記憶装置に置かれ、通常的方式でパリティが維持される。この代替実施例は、単一の記憶装置の故障後もパリティ・データが維持され続けるので、追加の障害保護レベルを実現する。しかし、予備記憶域のための追加の記憶空間が必要となることがあり、そうした予備記憶域が一時データ記憶など他の目的に通常使用される場合には、性能の低下を招くこともある。

【0016】第2の代替実施例では、記憶管理機構がホスト・システムのオペレーティング・ソフトウェアに常駐するが、その他の点では記憶制御装置に常駐する記憶管理機構と同じ機能を果たす。この実施例は、一般に好ましい実施例よりも速度が遅いが、記憶制御装置のコス

6

トを減少させることができる。

【0017】

【実施例】図1に、本発明の好ましい実施例のコンピュータ・システム100の主要構成要素の構成図を示す。ホスト・システム101が、バス102を介して記憶制御装置103と連絡する。記憶制御装置103は、プログラム式プロセッサ104、持久RAM (NVRAM) 105、排他的ORハードウェア (XOR) 108、キャッシュ・メモリ (RAM) 109を含む。持久RAM 105は、状況マップ106と目録 (TOC) 107を含む。制御装置103は、記憶装置121~124の動作を制御する。好ましい実施例では、記憶装置121~124は回転式磁気ディスク記憶装置である。図1には4台の記憶装置が示してあるが、記憶制御装置103に接続される実際の装置の数は変わり得ることを了解されたい。また、複数の記憶制御装置103をホスト・システム101に接続できることも了解されたい。好ましい実施例では、コンピュータ・システム100はIBMAS/400コンピュータ・システムであるが、どんなコンピュータ・システムも使用できる。

【0018】各記憶装置の記憶域は、ブロック131~138に分割されている。好ましい実施例では、すべての記憶装置の記憶容量が同じであり、すべてのパリティ保護されたブロックのサイズが同じである。本発明は様々なサイズの記憶装置または様々なサイズのブロックの構成で使用することも可能であるが、この好ましい実施例では制御機構が簡単である。

【0019】複数の記憶装置上の同じ位置にあるすべてのブロックのセットがストライプを構成する。図1で、記憶ブロック131~134が第1のストライプを構成し、ブロック135~138が第2のストライプを構成する。各ストライプ中のブロックのうちの1つがパリティ・ブロックに指定される。図1では、パリティ・ブロック131、136を斜線をつけて示してある。斜線をつけてない残りのブロック132~135、137~138は、データを記憶するためのデータ記憶ブロックである。ブロック131~134からなる第1のストライプのパリティ・ブロックが、ブロック131である。このパリティ・ブロックは、同じストライプ上の残りのブロック内のデータの排他的ORを含む。

【0020】好ましい実施例では、図1に示すようにパリティ・ブロックを異なる記憶装置間にラウンド・ロビン方式で分配する。書き込み動作ごとに、システムが同じストライプに書き込まれるデータを含むブロックだけでなくその同じストライプ用のパリティ・ブロックをも更新しなければならないので、通常はパリティ・ブロックの方がデータ・ブロックよりも頻繁に修正される。パリティ・ブロックを異なる記憶装置の間に分配すると、大部分の場合、アクセス作業負荷の分配により性能が向上する。ただし、本発明を実施する際にこのような分配は

(5)

特開平4-233025

7

必ずしも必要ではなく、代替実施例では単一の記憶装置上にすべてのパリティ・ブロックを置くことが可能である。

【0021】好ましい実施例では、各ストライプの1ブロックをパリティ情報専用に充てる。ある代替実施例では、ストライプのうちの1つはパリティ保護を含まない。このストライプは、保護を必要としない一時データ用に留保される。図8に、ブロック811～814からなるストライプにおけるこの代替実施例を示す。このブロックは、このパリティ・データ保護方式の一部分では

10 ない余分の記憶空間なので、任意のサイズでよい。
【0022】記憶域をそれぞれがデータ・ブロックとパリティ・ブロックを含むストライプに割り振る上記の方式は、上記の米国特許第4761785号明細書に記載されている割振りと同じである。

【0023】記憶制御装置103は、記憶管理プログラムを実行するプログラム式プロセッサ104を含む。記憶管理プログラムの動作については後で説明する。記憶制御装置103はまた、持久RAM105またはキャッシュRAM109内のデータの排他的ORを計算する排

20 他的ORハードウェア108を含む。ある代替実施例では、プロセッサ104で排他的OR演算を行うことができるが、この目的用の特別のハードウェアを使うと性能が向上する。
【0024】持久RAM105は、記憶装置に物理的に書き込まれるのを待つデータ用の一時待機域として記憶制御装置103が使用する。持久RAM105には、この一時データに加えて、状況マップ106と目録(TOC)107が記憶される。目録107は、書き込まれるのを待っているデータの、記憶装置内のそれが記憶され

30 位置へのマッピングを含んでいる。
【0025】状況マップ106は、各データ・ブロック用の対応するパリティ・ブロックの位置、および障害回復モード中の各データ・ブロックの状況を識別するために使用される。状況マップ106は、図2に詳しく示してある。状況マップ106は、各記憶装置ごとの状況マップ記入項目のテーブルを含んでいる。各状況マップ記入項目201は、記憶装置上のデータ・ブロックの位置202、障害モードで動作しているときそのデータを回復する必要があるか否かを示す状況ビット203、およ

40 び対応するパリティ・ブロックの位置204を含む。
【0026】再度図1を参照すると、キャッシュ・メモリ109は、非持久RAMであり、記憶装置から読み取ったデータを記憶するのに使用される。キャッシュ・メモリ109は、読取り動作時に記憶装置からホスト・システム101にデータを転送する際に、バッファとして働く。その上、そのデータの修正および書直しの確率が高いとホスト・システム101から指示があったときも、それに応答して、データがキャッシュ109にセーブされる。未修正のデータは、対応するパリティ・デー

8

タの更新のために、修正済みデータと排他的ORしなければならないので、読み取ったデータをキャッシュ109にセーブすると、書込み動作の直前にそれを再度読み取る必要をなくすることができる。キャッシュ109が存在するのは、性能を向上させるためにすぎない。ある代替実施例では、それなしに本発明を実施することも可能である。記憶装置から読み取られたデータを持続メモリに保存することはシステムの保全性にとって必要でない

ので、キャッシュ109は非持久RAMとして特定してある。しかし、キャッシュは持久RAM105の一部として実施することもできる。メモリ・モジュールの相対的コストとサイズに応じて、そのような手法が望ましいことがあり得る。
【0027】本発明に必要なハードウェアおよびソフトウェアの諸特徴に関するシステムの機能について以下で述べる。システムは、正常モードと障害モードの2つの動作モードを有する。すべてのディスク記憶装置が適正に機能しているとき、システムは正常モードで動作する。1台の記憶装置が故障したとき、動作モードは障害

モードに変わるが、システムは動作し続ける。
【0028】正常モードでの読取り動作を図3に示す。読取り動作を実行するには、ステップ301で、ホストからのREAD(読取り)コマンドを受け入れ、ステップ302で、要求されたデータが持久RAM105またはキャッシュ109内に存在するかどうか判定する。存在する場合は、ステップ304で、持久RAMまたはキャッシュ内のデータがホストに直接送られる。そうでない場合は、ステップ303で、まず適当な記憶装置からキャッシュ109にデータが読み込まれ、次にステップ304で、そこからホスト・システムに転送される。キャッシュ109は、書込み動作中にも性能の改善をもたらす。書込み動作が処理されるときに、更新すべきデータの元のバージョンが既にキャッシュ109内にある場合、パリティを更新するために再度データを読み取る必要はなく、したがってシステム性能が改善される。キャッシュ109の内容は、当技術分野で既知の様々なキャッシュ管理技術のいずれかをを用いて管理する。

【0029】書込み動作は、記憶制御装置のプロセッサ104内で走行する2つの非同期タスクによって実行される。1つのタスクは、バス102を介してホストと通信するものである。図4にこれを示す。ステップ401でホストからWRITE(書込み)コマンドを受け入れたとき、書込み動作が開始する。次にステップ402で、目録107を検査して、記憶装置に書き込むべきデータを記憶するのに十分な空間が持久RAM105内で利用可能かどうか判定する(利用可能な空間には、書き込むべきデータのバック・レベル・バージョンが使用する空間と、未使用の空間が含まれる)。空間が利用可能でない場合、制御装置103はホストからデータを受け

50 取ることができず、ステップ403で空間が利用可能と

(6)

特開平4-233025

9

なるのを待たなければならない(すなわち、持久RAM 105内に既にあるデータが記憶域121~124に書き込まれるのを待たなければならない)。持久RAM 105内で空間が利用可能になったとき、ステップ404でホスト101から持久RAM 105にデータがコピーされ、目録107が更新される。次いでステップ405で、プロセッサ104がホストに動作完了メッセージを発行する。動作完了メッセージを受け取った後、ホストは、あたかも記憶域121~124に実際にデータが書き込まれたかのように、自由に処理を続けることができるが、実際にはデータが持久RAM 105中で暫らく待つこともある。ホストから見ると、この動作は完了しているように見える。

【0030】第2の非同期タスクは、持久RAM 105から記憶装置にデータを書き込むものである。正常モードでのこのタスクの流れ図を図5に示す。このタスクは、ステップ501で、持久RAM 105内で待機している書き込み動作のうちからある書き込み動作を選択する。この選択基準は本発明の一部ではなく、たとえば先入れ先出し、後入れ先出し、あるいはシステム性能その他の考慮点に基づく他の何らかの基準でよい。書き込み動作が実行されるとき、パリティを更新しなければならない。新しい書き込みデータと古いデータの排他的ORを取ることにより、その書き込み動作によって変更されるビットのビット・マップを得ることが可能である。このビット・マップを既存のパリティ・データと排他的ORすると、更新されたパリティ・データが得られる。したがって、記憶装置に書き込む前に、まずステップ502で、古いデータが未修正の形でキャッシュ109内に存在するかどうかを検査する。存在しない場合は、ステップ503で、そのデータをキャッシュ109に読み込む。キャッシュ109内の古いデータが、ステップ504で、持久RAM 105内の新しいデータと排他的ORされ、変更済みデータのビット・マップを生成する。このビット・マップは持久RAM 105に一時的にセーブされ、新しいデータは記憶装置121~124の1つに書き込まれる。次にステップ506、507で古いパリティ・データがキャッシュ109に読み込まれ(まだそこない場合)、ステップ508で、それがビット・マップと排他的ORされて新しいパリティ・データを生成する。ステップ509で、この新しいパリティ・データが、記憶装置121~124の1つに書き込まれ、目録107が更新されて、書き込み動作は完了する。

【0031】記憶装置の故障が検出されると、システムは障害モードで動作し始める。記憶装置の故障とは、それが機能できない、すなわちデータにアクセスできないという意味である。このような故障は、必ずしも装置自体の破壊によって起こるとは限らない。たとえば、装置の電源が切れることもあり、データ・ケーブルが切断されることがあり得る。システムから見ると、原因が何で

10

あろうと、そのようなどんな故障も記憶装置の故障である。そのような故障を検出する検出機構は、当技術分野で知られている。一般的な機構としては、応答を受け取る前にタイムアウトになることや、受け取ったデータで高いエラー率が続くことがある。

【0032】図6に、システムが障害モードで動作している際の読取り動作を示す。正常モードの読取り動作の場合と同様に、ステップ601でホストからの読取り動作が受け入れられたとき、ステップ602で、制御装置はまず、所望のデータがあるかどうかその持久RAM 105およびその非持久キャッシュ109を検査する。持久RAMまたはキャッシュ内にデータが存在する場合、データがシステム・バス102を介してホストに転送される。持久RAMまたはキャッシュ内にデータがなく、故障していない記憶装置上にある(ステップ603)場合は、ステップ604で、記憶装置からキャッシュ109に通常の方式でそのデータが読み込まれる。データが故障した記憶装置上にある場合、ステップ605で、制御装置は、状況106内の状況マップ記入項目201を検査して記憶装置内での所期のデータの位置を調べる。状況マップ記入項目は、データが回復されたかどうか、すなわち排他的ORによってデータが再構築されて、ある代替位置に記憶されたかどうかを示す。状況マップが、データが回復されていないことを示す(ステップ605)場合、ステップ608で、制御装置は、故障した装置以外のすべての記憶装置上の対応する位置を次々に読み取る。読み取られた各データ・ブロックは、XORハードウェア108によって、以前に読み取られたブロックのXOR結果の累計とXORされる。最終的なXOR結果が、故障した装置の再構築されたデータとなる。ステップ609で、この再構築データが、このデータ・ブロックに対応するパリティ・ブロックに書き込まれる。このブロックの位置が、状況マップ108のパリティ・ブロック・アドレス・フィールド204に記憶される。回復されたデータがパリティ・ブロック位置に書き込まれた後、ステップ610で、同じストライプ中の各ブロックの状況ビット203を、データが回復されたことを示す“1”に変更することにより、状況マップ108が更新される。ステップ611で、再構築データがホストに送られる。状況ビット203が、最初から、データが回復されたことを示す“1”を含む場合、ステップ606で、制御装置は、状況マップから以前のパリティ・ブロック域の(回復されたデータが記憶される)位置を得て、ステップ607で、この位置からのデータをキャッシュ109に直接読み込むことになる。この装置により、特定のデータ・ブロックを読み取るのに、すべてのディスク記憶装置を1回読み取るだけでよくなる。データが回復されると、そのデータの物理的記憶位置が、パリティ記憶域用に以前に使用されていた位置に有効に再配置され、その後はそのブロックを読み取るのに、そ

(7)

特開平4-233025

11

の1台の記憶装置を読み取るだけでよくなる。

【0033】図7に、システムが障害モードで動作しているときの記憶装置への書き込み動作を示す。正常モードの書き込みの場合と同様に、図4に示したホスト通信タスクが、書き込むべきデータをホストからバス102を介して受け取る。記憶装置書き込みタスクは、ステップ701で、持久RAM105内の待ち行列からある書き込み動作を選択する。制御装置が、故障した装置にデータを書き込むかどうか判定し（ステップ702）、状況マップを検査する（ステップ703、709）。故障した装置にデータを書き込むのだが、そのブロック内のデータがまだ回復されていない場合は、そのブロックを回復してからでないと書き込み動作は可能にならない。回復は、読取り動作について上述したのと同じステップに従う。ステップ704で、同じブロック・ストライプ中の（パリティ・ブロックを含む）各ブロックが読み取られ、その内容が、以前に読み取ったブロックの排他的ORの累計と排他的ORされる。その結果が再構築されたデータであり、ステップ705で、それがそのパリティ・ブロックに使用される位置に書き込まれる。そのブロック全体の回復が完了すると、ステップ706で、新しいデータ（通常はそのブロックの一部分だけに及ぶ）が、以前のパリティ位置で回復されたデータの上に書き込まれ、ステップ707で、そのブロックが回復されたことを示すように状況マップが更新される。故障した装置にデータを書き込むのだが、データがすでに回復されている場合には、ステップ708で、それが、今は回復されたデータを記憶するのに使用されている、以前のパリティ位置に直接書き込まれる。

【0034】障害モードで動作しているときに、故障していない装置にデータが書き込まれつつある場合、ステップ709で、制御装置は状況マップを検査する。状況が“1”で、故障した装置上の同じストライプ中のデータ・ブロックがすでに回復されていることを示す場合、ステップ710で、書き込みデータがその故障していない記憶装置に直接書き込まれる。状況が“0”の場合は、故障していない装置に直接データを書き込むことはできない。というのは、そのような動作はパリティを変更し、後で故障した記憶装置内の対応するデータを再構築するのが不可能になるからである。したがって、この好ましい実施例では、制御装置がまず、故障した装置上の同じストライプ中のデータ・ブロックを回復する。図7に示すように、ステップ711で、故障した記憶装置内のデータ・ブロックがまず排他的ORによって再構築され、ステップ712で、上記のステップに従ってパリティ・ブロック位置にセーブされる。次いでステップ713で、書き込みデータがその記憶装置に書き込まれ、ステップ714で状況マップが更新される。書き込むべきデータを含むストライプのパリティ・ブロックが故障した装置上にある場合は、パリティが何らかの形で失われる

12

ので、再構築は不要である。したがって、記憶装置の故障が検出されたとき、このストライプ上のすべてのブロックの状況が“1”にセットされる。その効果は、このストライプ上のデータを、故障した装置上の対応するブロックが既に回復されている場合と同様に、記憶装置に直接書き込ませることである。たとえば図1を参照すると、記憶装置121が故障した場合、制御装置は直ちにブロック132～134の状況を“1”にセットして、これらのブロックへの書き込み動作が直接進行できるようにする。ある代替実施例では、書き込み動作が、故障していない装置に対するものであり、故障している装置上の対応するブロックが回復されていない場合、正常モードの書き込み動作に使用するのと同じステップに従ってパリティ・ブロックを更新し、故障した装置上のデータの読取りまたは書き込みが要求される場合に、後で故障した装置のデータを再構築する能力を保存することが可能となる。

【0035】好ましい実施例では、パリティ・ブロックを使って再構築されたデータを記憶し、その結果、1台の記憶装置が故障した後にシステムはパリティ保護なしで走行することになる。図8に示すように、記憶装置上に十分に大きな予備記憶ストライプを留保しておく、代替実施例も可能である。この予備記憶ストライプは、パリティ保護を必要とせず、必要が生じた場合に重ね書きできる一時データを含むことができ、また全くデータを含まないこともできる。この代替実施例では、再構築されたデータが、パリティ・ブロックの代りに予備記憶ストライプ811～814のブロックに再配置される。この代替実施例が可能なのは、故障した装置の非予備内容を収容するのに十分な予備記憶域が存在する場合だけである。これもまた、システムが利用できる一時記憶域の量を減少させる結果を招き、性能が低下したり、システムがサービスできるユーザの数が減少したりする可能性がある。この代替実施例では、正常モードの読取り動作と書き込み動作は、好ましい実施例と全く同様に行われる。障害モードで動作しているとき、状況マップが検査され、必要に応じて上記のようにしてデータが再構築される。しかし、再構築されたデータは、パリティ・ブロックには書き込まれず、予備記憶域中のブロックに書き込まれる。状況マップ106中に、故障した装置に含まれていたデータの新しい位置を記憶する別のフィールドが必要である。さらに、書き込み動作の際に、正常モードの書き込み動作の場合と同様にしてパリティが更新される。これは、故障した装置上のデータが再構築された後に行われる。

【0036】もう1つの代替実施例では、パリティ保護とミラー構成が同じシステム上で組み合わせられる。記憶装置上に含まれるデータのあるものは本明細書に記述したパリティ保護機構によって保護され、別のデータはミラー構成される。記憶装置が故障した場合、パリティ保

(8)

特開平4-233025

13
隠されたデータは上述のように再構築されて記憶され、ミラー構成のデータは、ミラー・コピーを含む記憶装置からアクセスされる。

【0037】
【発明の効果】本発明により、多数のデータ記憶装置をもつコンピュータ・システムが、データ喪失から回復でき、データ記憶装置が1台故障しても引き続き動作できる、改良された方法および装置が提供される。

【図面の簡単な説明】

【図1】本発明の好ましい実施例の構成要素を組み込んだシステムの構成図である。

【図2】状況マップを示す図である。

【図3】正常動作モード中の読取り動作に関する諸ステップの流れ図である。

【図4】書き込むべきデータをホストから記憶制御装置に転送する際に関する諸ステップの流れ図である。

【図5】正常動作モードで記憶装置にデータを書き込む

14
際に関する諸ステップの流れ図である。

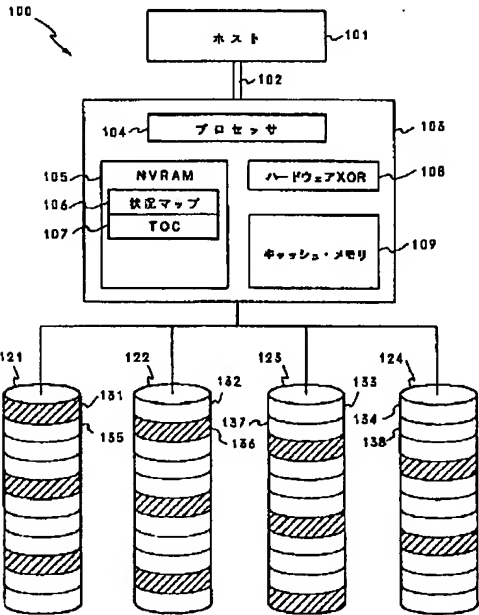
【図6】記憶装置故障後の読取り動作に関する諸ステップの流れ図である。

【図7】ある記憶装置が故障したとき、記憶装置にデータを書き込む際に関する諸ステップの流れ図である。

【図8】本発明の代替実施例による構成要素を組み込んだシステムの構成図である。

- 【符号の説明】
- 100 コンピュータ・システム
 - 101 ホスト・システム
 - 103 記憶制御装置
 - 104 プログラム式プロセッサ
 - 105 持久RAM
 - 106 状況マップ
 - 107 目録（TOC）
 - 108 排他的ORハードウェア
 - 109 キャッシュ・メモリ

【図1】



【図2】

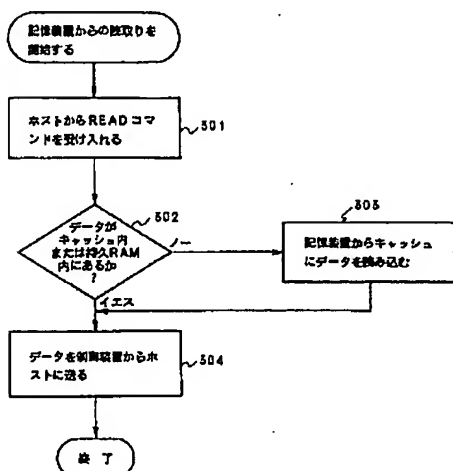
202	203	204
ブロック・アドレス	状況ビット	パリティ・ブロック・アドレス
131	1	131
135	0	136
...
132	1	131
136	0	136
...
133	1	131
137	0	136
...

104

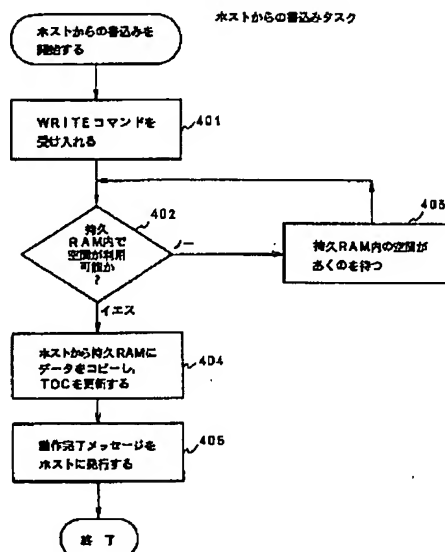
(9)

特開平4-233025

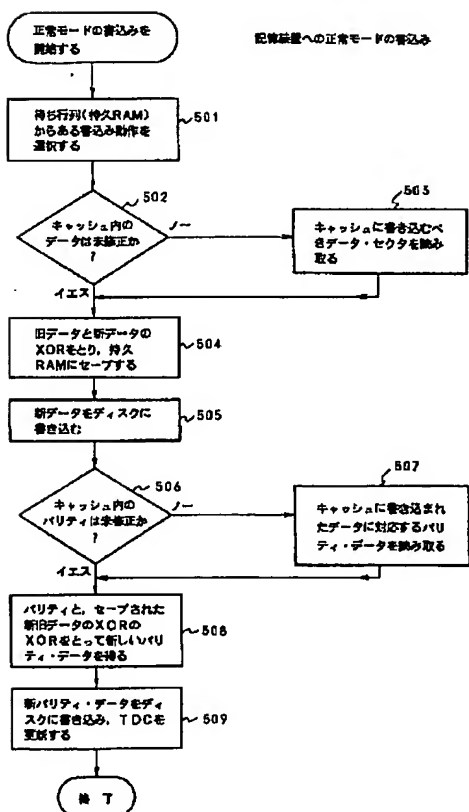
【図3】



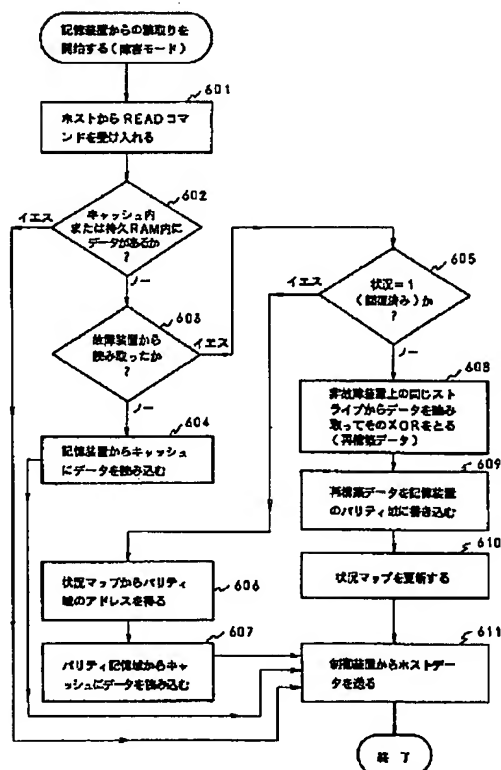
【図4】



【図5】



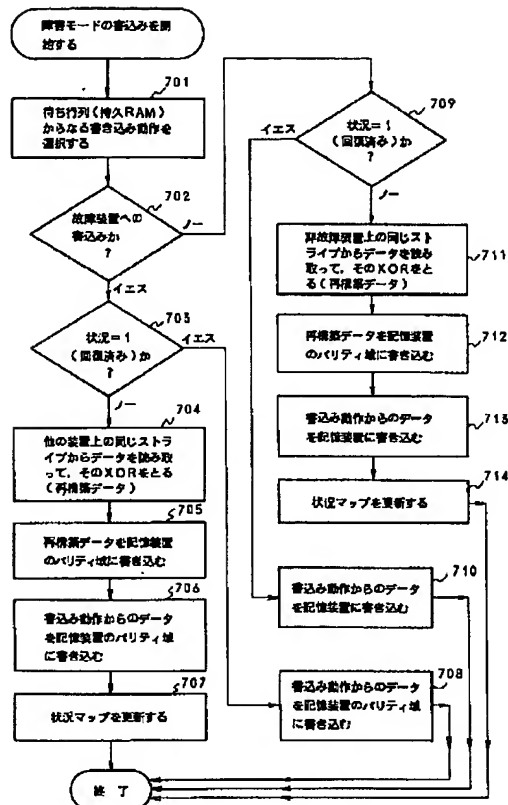
【図6】



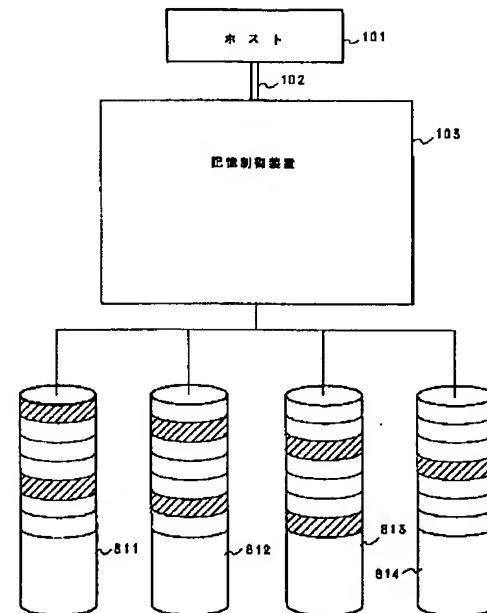
(10)

特開平4-233025

【図7】



【図8】



フロントページの続き

(72)発明者 プリアン・エルドリッジ・クラーク
 アメリカ合衆国 55904、ミネソタ州、ロ
 チエスター、ウッドパイン・コート・サウ
 ス・イースト 6810番地

(72)発明者 レイモンド・スペンサー・マツクロバーツ
 アメリカ合衆国 55901、ミネソタ州、ロ
 チエスター、オーク・ノール・レーン・ノ
 ース・ウエスト 1907番地